

Data Augmentation and Reversible Jump MCMC for Multinomial Index Problems

Matthew R. Schofield^{1*} and Richard J. Barker²

¹Department of Statistics, Columbia University, New York, NY, USA.

²Department of Mathematics and Statistics, University of Otago,

P.O. Box 56, Dunedin, New Zealand.

Abstract

A feature of multinomial models with unknown index N is that the dimension of the parameter space potentially depends on N , a complication when fitting models by Markov chain Monte Carlo (MCMC). Two commonly used approaches to this problem are: (i) trans-dimensional reversible jump MCMC and (ii) superpopulation data augmentation. A distinguishing feature of the two approaches is that N , and combinatorial terms involving N , are not explicit in the superpopulation likelihood. To resolve ambiguity about the relationship between the two approaches we compare them analytically. We show that superpopulation data augmentation is equivalent to trans-dimensional sampling but with a restricted prior on N . We highlight potential drawbacks that result from not making N explicit in the likelihood in the superpopulation approach. One advantage of the superpopulation approach has been the availability of easy to use BUGS code. We provide simple BUGS code that implements trans-dimensional reversible jump MCMC for the mark-recapture model M_h that can be readily extended

*Current Address: Department of Statistics, University of Kentucky, Lexington, KY, USA. E-mail: matthew.schofield@uky.edu

to related models.

Key Words: BUGS/JAGS; Capture-recapture; Data augmentation; Heterogeneity; Multinomial; Reversible jump MCMC.

1 Introduction

Inference about N , the unknown size of a population, based on capture-recapture models is of enormous interest, with ecology and epidemiology just two areas in which there are many applications. This is a specific case of a more general class of problems in which the index of a multinomial model is unknown. A sub-class of models of particular interest are the heterogeneity models in which capture probabilities vary among individuals. Link (2003) has shown that inference about N within this class of models is sensitive to the choice of model for describing individual variation.

A drawback of much available software for fitting M_h is that analyses are restricted to default choices of model for the individual capture probabilities. Restricting analyses based on available computer code seems undesirable and an advantage of Bayesian approaches is that hierarchical modelling and data augmentation implemented via Markov chain Monte Carlo (MCMC) provide a natural framework for general model fitting.

A commonly used dataset for illustrating model M_h is the snowshoe hare data provided by Otis et al. (1978) in which 68 individuals were captured in six days of trapping with encounter frequencies $f = (25, 22, 13, 5, 1, 2)'$ where f_j ($i = 1, \dots, k$) is the number of hares caught i times during the $k = 6$ study. Define N as the population size, y_i as the number of times that individual i was caught and p_i as the corresponding capture probability. To model these using a program such as BUGS (Lunn et al. 2000) or JAGS (Plummer 2003) we would like to be able to specify the model simply as

$$y_i \sim \text{Bin}(k, p_i)$$

for i in $1 : N$ (herein, unless we specify the two programs separately, we will use

BUGS to refer to BUGS and JAGS jointly). To complete the model specification we would then specify a suitable distribution for p_i , hyperpriors for any parameters of this distribution and a prior for N . However two related things prevent us from doing this: (1) BUGS does not allow the index of the loop to be stochastic and (2) the dimension of p changes each time we update N to a new value.

Royle et al. (2007) provide a neat solution that is easily implemented in BUGS based on a superpopulation. However, a drawback of their model and code is that N does not appear in the likelihood; rather it is a derived parameter. This makes specification of priors on N more of a challenge than if N appears explicitly in the likelihood.

Here we contrast approaches to fitting M_h using the models of King and Brooks (2008) and Royle et al. (2007) when using Bayesian methods. The approach taken by King and Brooks (2008), as well as Durban and Elston (2005), Schofield and Barker (2008) and Wright et al. (2009) for similar models, is to consider a trans-dimensional (TD) algorithm (Carlin and Chib 1995, Green 1995) such as the reversible jump MCMC (RJMCMC) algorithm (Green 1995). In contrast, Royle et al. (2007) as well as Royle and Dorazio (2008), Link and Barker (2010), Schofield and Barker (2010) use a data augmentation (DA) approach (Tanner and Wong 1987) to fit the model.

Our purpose is to:

1. Clarify the distinction between the different likelihoods used to fit M_h models by MCMC.
2. Provide simple BUGS code for M_h that explicitly includes N as a parameter.
3. Consider the comment made by Royle et al. (2007) that their approach is not a case of trans-dimensional MCMC and fundamentally differs from approaches that use RJMCMC.

2 M_h Likelihoods

Likelihoods for model M_h include an integrated likelihood (Burnham and Overton 1978, Otis et al. 1978), a “full” likelihood (Pledger 2000, King and Brooks 2008), or a super-population likelihood (Royle et al. 2007). The three most obvious differences between these approaches are (i) whether or not N is explicitly included in the likelihood, (ii) the presence or absence of combinatorial terms involving N , and (iii) the presence or absence of a superpopulation (with corresponding value M) in the model. The likelihood of Royle et al. (2007) does not include N , or combinatorial terms involving N , but does include M . The full likelihood approach of Pledger (2000) explicitly includes N and associated combinatorial terms, does not include M and treats the vector of capture probabilities \mathbf{p} as parameters. Burnham and Overton (1978) explicitly includes N , does not include M and models \mathbf{p} as random effects; combinatorial terms appear once the capture probabilities are integrated out of the likelihood.

Table 1 about here

2.1 Burnham’s Integrated Likelihood

A complete data model can be written as $\Pr(\mathbf{X} = \mathbf{x} | \mathbf{P} = \mathbf{p})$; as shorthand we use $[\mathbf{x} | \mathbf{p}]$. Both \mathbf{x} and \mathbf{p} are unobserved, although we know values of some rows of \mathbf{x} through the observed data matrix \mathbf{x}^{obs} . Following Burnham and Overton (1978) we can write the model as

$$[\mathbf{x} | \mathbf{p}] = \prod_{i=1}^N \prod_{j=1}^k p_i^{x_{ij}} (1 - p_i)^{1-x_{ij}} \quad (1)$$

under the assumption that, conditional on $\mathbf{P} = \mathbf{p}$, capture events are independent among individuals and occasions. Conditional on N , which is treated as a parameter, the number of rows of \mathbf{x} are now known.

As noted by Burnham and Overton (1978) this model is over-parameterized and not useful for estimation. The usual remedy for over-parameterization is to model \mathbf{P} as

random effects sampled from a distribution with *pdf* $f_{\mathbf{P}}(\mathbf{p}|\theta_p)$ for $\mathbf{P} \in \mathbb{R}_{[0,1]}^N$. We then integrate over \mathbf{p} to obtain the integrated likelihood given by Burnham and Overton (1978); this is the observed data likelihood (Gelman et al. 2004).

2.2 Complete Data Likelihood (CDL)

The model (1) is underspecified. Adding the random effects distribution for \mathbf{P} we can write a CDL as

$$[\mathbf{x}|\mathbf{p}][\mathbf{p}|\theta_p] = [\mathbf{p}|\theta_p] \times \prod_{i=1}^N \prod_{j=1}^k p_i^{x_{ij}} (1 - p_i)^{1-x_{ij}}. \quad (2)$$

As before, both \mathbf{x} and \mathbf{p} are unobserved. Associated with \mathbf{x}^{obs} is a subspace of \mathcal{S} , which we denote by $\Omega(\mathbf{x})$, of capture history matrices that are identical in appearance to \mathbf{x}^{obs} . The only differences among elements of $\Omega(\mathbf{x})$ are that the row indices are permuted.

Under the assumption that $x_i|p_i$ and $x_j|p_j$ are independent for all $i \neq j$, it follows that the x 's are exchangeable given the p 's and that the elements in $\Omega(\mathbf{x})$ have identical joint probabilities across their rows. A heuristic explanation is that once we have drawn (x_i, p_i) pairs from their joint distribution they are linked – when we permute the x 's the p 's get permuted with them.

It now follows that

$$\begin{aligned} [N, \mathbf{p}|\mathbf{x}^{obs}, \theta_p] &\propto [\mathbf{x}^{obs}|\mathbf{p}][\mathbf{p}|\theta_p] \\ &= \frac{N!}{\prod_{h \in \mathcal{H}} z_h!} \times \prod_{i=1}^N [p_i|\theta_p] \prod_{j=1}^k p_i^{x_{ij}} (1 - p_i)^{1-x_{ij}} \end{aligned} \quad (3)$$

is also a likelihood where $\frac{N!}{\prod_{h \in \mathcal{H}} z_h!}$ is the dimension of $\Omega(\mathbf{x})$. The likelihood (3) is not the observed data likelihood (ODL) as it still contains unobserved elements. If we integrate across the random effects we obtain the ODL given by Burnham and Overton (1978). However, the likelihood (3) is useful in Bayesian inference carried out by MCMC in that explicit modelling in terms of latent components can be a useful device for improving

mixing.

2.3 Super Population CDL

Royle et al. (2007) describe a CDL in terms of a super-population M (essentially an upper bound on the population size N), with \mathbf{X} now an M by k matrix. The idea is that we replace N in the likelihood with the outcomes \mathbf{w} of the random vector \mathbf{W} of dimension M where W_i takes the value 1 if individual i is included in the population and 0 otherwise such that $N = \sum_{i=1}^M W_i$. By assumption, \mathbf{W} and \mathbf{P} are independent given their parameters. The complete model on which the Royle et al. (2007) approach is based can be written as

$$\begin{aligned}
[\mathbf{x}, \mathbf{w}, \mathbf{p} | \alpha_p, \psi] &= [\mathbf{w} | \psi][\mathbf{p} | \mathbf{w}, \alpha_p][\mathbf{x} | \mathbf{w}, \mathbf{p}] \\
&= [\mathbf{w} | \psi][\mathbf{p} | \alpha_p][\mathbf{x} | \mathbf{w}, \mathbf{p}] \\
&= \prod_{i=1}^M \prod_{j=1}^t (p_i w_i)^{x_{ij}} (1 - p_i w_i)^{1-x_{ij}} f_P(p_i) \psi^{w_i} (1 - \psi)^{1-w_i} \\
&= \prod_{i=1}^M (p_i w_i)^{y_i} (1 - p_i w_i)^{t-y_i} f_P(p_i) \psi^{w_i} (1 - \psi)^{1-w_i}. \tag{4}
\end{aligned}$$

The sample space \mathcal{S} for \mathbf{x} is now of dimension 2^{Mt} and the sample space \mathcal{W} for \mathbf{w} of dimension 2^M . The sample space for \mathbf{p} is given by $\mathcal{P} = \mathbb{R}_{[0,1]}^M$.

Denoting the *cdf* for p by $F(p) = F_{\theta_p}(p)$ and integrating over \mathbf{p} and \mathbf{w} , we obtain

$$[\mathbf{x}^{obs}, N | \alpha_p, \psi, M] = \frac{N!}{\prod_{h \neq 0} z_h! (N - n)!} \prod_{j=0}^t \theta_{jF}^{f_j} \times \binom{M}{N} \psi^N (1 - \psi)^{M-N},$$

where $\theta_{jF} = \int_0^1 p^j (1 - p)^{t-j} dF(p)$. That is, we have Burnham's marginal likelihood multiplied by the prior for N induced by the model $W_i \sim \text{Bern}(\psi)$.

Inference in the two approaches will be identical for the appropriate choice of prior on ψ in the Royle et al. (2007) likelihood. For example,

$$f(\psi) \propto 1$$

corresponds to a discrete uniform prior on N as shown by Royle et al. (2007). Another choice could be:

$$f(\psi) \propto \frac{1}{\psi}$$

which leads to $N \propto \frac{1}{N}$ but with the upper bound M . This is Jeffrey’s prior for N but truncated to the range $[0, M]$. In general, the prior on N induced in the super-population approach is the marginal distribution for N arising from a binomial mixture, with mixing distribution given by the prior on ψ . While it is relatively easy to construct reference priors of the form given above, constructing general priors on N that are not binomial mixtures will be more difficult. We do not wish to overstate this problem as the binomial mixtures provide a flexible class of priors including the beta-binomial and the negative-binomial as a limiting case.

3 Gibbs Sampling and Model M_h

Durban and Elston (2005), King and Brooks (2008) and Schofield and Barker (2008) all used a trans-dimensional (RJMCMC) approach to fit M_h using the CDL described above. Wright et al. (2009) used a similar approach in an application where individual identities were uncertain. In this approach, sampling from the full conditional for N includes an explicit reversible-jump step in the sense of Green (1995). In contrast, the reversible-jump step is unnecessary in the super-population approach of Royle et al. (2007).

To describe what we call the “standard approach” to Gibbs sampling we start with the complete data likelihood (2). Although we describe model fitting from a Bayesian stand-point, we could implement a Frequentist approach using an EM-algorithm instead of Gibbs sampling.

For simplicity we assume θ_p is known. A Gibbs sampler can be constructed by alternating sampling from the following distributions,

1. The full conditional distribution for $p_i, \forall i$:

$$[p_i|x, N, \theta_p] \propto p_i^{x_{ij}} (1 - p_i)^{1-x_{ij}} f(p|\theta_p).$$

Depending on $[p|\theta_p]$, we can either sample from this distribution directly, or use another a simulation based sampler, such as the Metropolis-Hastings algorithm or rejection sampling.

2. The full conditional distribution for N :

$$[N|x, p] \propto \frac{N!}{(N-n)!} \prod_{i=1}^N \prod_{j=1}^k p_i^{x_{ij}} (1 - p_i)^{1-x_{ij}} f(N).$$

This update is more difficult since the value of N changes the dimension of p , so we use a TD algorithm. Following Schofield and Barker (2008) we use a special case of the RJMCMC algorithm. We first propose a new value N^* from $J(N^*|N)$ and specify p^* to be a vector of length N^* corresponding to the capture probabilities for the N^* individuals. We set $p_i^* = p_i$ for all $i = 1, \dots, N$ and generate p_i^* for $i = N+1, \dots, N^*$ from the prior distribution $[p|\theta_p]$. Generating p^* in such a way simplifies the RJMCMC algorithm. The acceptance probability is

$$q = \frac{N^*!(N-n)!}{(N^*-n)!N!} \prod_{i=N+1}^{N^*} (1 - p_i^*)^k \frac{J(N|N^*)}{J(N^*|N)} \frac{f(N^*)}{f(N)}.$$

If our proposed value $N^* < N$ we obtain p^* by removing the last $N - N^*$ rows of p . The acceptance probability becomes

$$q = \frac{N^*!(N-n)!}{(N^*-n)!N!} \frac{1}{\prod_{i=N^*+1}^N (1 - p_i)^k} \frac{J(N|N^*)}{J(N^*|N)} \frac{f(N^*)}{f(N)} I(N^* \geq n),$$

where $I(\cdot)$ is an indicator function. Importantly, we note that the full RJMCMC expression has been reduced to a standard Metropolis-Hastings like expression that consists of a likelihood ratio \times jumping distribution ratio \times prior ratio.

If we also wish to estimate θ_p , we simply need to (i) specify a prior distribution for

θ_p , and (ii) sample from the full conditional distribution for θ_p in turn. Although we have never found it necessary, more complicated RJMCMC setups could be specified in order to speed up the RJMCMC algorithm. For more extensive details of RJMCMC we direct the reader to Gelman et al. (2004).

A strength of this approach is the ease with which we can incorporate extensions to the model. For example, if the tags are themselves uncertain we need to make only one adjustment to the algorithm above: we add a step where we sample from the full conditional distribution for all of the uncertain tags (see Wright et al. 2009 for details). We do not need to alter the steps we already have because conditional on x the terms $y_h!$ are fixed and cancel out of the above distributions. With N explicit in the model as a parameter it is also easy to include different priors for N , including hierarchical priors, in the usual way. For example, if mark-recapture is carried out across distinct subpopulations it would be straightforward to model N for each population in terms of an appropriate spatial model.

One perceived weakness of this TD approach is that it is difficult to incorporate into BUGS, in contrast to the super population approach (Royle et al. 2007). However, this is not the case, so long as we specify a prior for N that has an upper bound M (Durban and Elston 2005). We provide BUGS code for model M_h (Figure 1) that is just as simple and requires similar programming effort to the code given by Royle et al. (2007). A full description of the model including the data and initial values files are available for both BUGS and JAGS at www.maramatanga.com. We also provide alternative model/data files for model M_h that allow for easier specification of the data, albeit with a longer model statement.

Note that specifying an upper limit is required *only* for easy implementation in BUGS; there is no superpopulation in the Royle et al. (2007) sense. We can achieve an approximation to any distributions for N with support on the non-negative integers by including a large M and truncating (JAGS) or censoring (BUGS) the distribution at M (although very large values for M will slow model fitting). We find that both (i) the time taken to run the BUGS code and (ii) the mixing of the MCMC output

is approximately equivalent to the corresponding superpopulation code. However, we note that many things can substantially influence both the time taken for the code to run and the mixing performance. These include (i) how we parameterize the model on N , (ii) the sampling algorithms the programs choose and (iii) which software package and version we use. We find that in general for models of this type, JAGS takes longer to run than BUGS but mixes better.

We illustrate the code in Figure 1 with model M_h using the snowshoe hare data. The model was fitted in JAGS 2.1.0 and density plots of N (Figure 2) can be compared to those in Royle et al. (2007) and Link and Barker (2010).

Figure 1 about here

Figure 2 about here

3.1 Super Population approach and RJMCMC

The starting point for the super population model is the complete data likelihood specified in (4). To complete the specification of the model we specify a prior for ψ , $f(\psi)$, as well as a random effects distribution for p . Royle et al. (2007) provide BUGS code for fitting this model. The CDL that their BUGS code provides, referred to as a data augmentation (DA) algorithm, is identical to that in (4) with the exception that Royle et al. (2007) fix the first n values of W_i at 1. Data augmentation now occurs over a reduced subspace for \mathbf{W} ; that in which $W_1 = W_2 = \dots = W_n = 1$. However, the effect on the reduced CDL is simply the omission of the $\prod_h z_h!$ terms. Therefore, the likelihood is equivalent, although as we discuss below a consequence of conditioning on this one order \mathbf{W} is that the observed z_h 's cannot be modelled in the presence of tag-reading uncertainty.

Royle et al. (2007) comment that the DA algorithm is an estimation approach which fundamentally differs from other TD algorithms which are “model-selection approaches” since the use of a TD algorithm treats each value N as a separate model.

Here we argue that this difference is semantic, and that in an algorithmic sense the two approaches are equivalent since they can be set up to give the same Gibbs sampler.

To see the equivalence, we start with a superpopulation model similar to the that of Royle et al. (2007). The only difference is that we treat the capture probability as undefined for any individual that is not in the population, that is, p_i is undefined when $w_i = 0$. It is clear that this model is TD, in the sense that changing the value of w_i changes the dimension of the capture probability vector \mathbf{p} whose length is defined by $N = \sum_i w_i$. This setup has 2^{M-n} possible “models” (by possible models we refer to models with non-zero probability). That is, N is censored by n and has an upper limit of M so that there are $M - n$ pseudo-individuals who were unobserved. Each of these 2^{M-n} models describes the different ways that the $M - n$ unobserved individuals could potentially be included or excluded from the population. To complete the TD specification we require a number of additional components. Using the language of Gelman et al. (2004) pg 338-339, we require

1. Auxiliary random variables that make the parameter space equivalent for each of these 2^{M-n} models. In particular, we need to ensure that we have M parameters (equal to the dimension of p if all M individuals are included in the population). For each of the 2^{M-n} models we need a distribution that describes how to generate the auxiliary random variables when moving to another model.
2. A set of bijections that describe how the set of parameters and auxiliary variables in one model relates to the set of parameters and auxiliary variables in another model.
3. Prior model probabilities.

We specify these components as follows:

1. Set the distribution for the auxiliary variables for each pair of models to be the prior distribution $f(p|\theta_p)$. This simplifies the algorithm since this distribution is the same for each of the auxiliary random variables in every model.
2. Specify identity bijections between each pair of models.

3. Set the prior model probability as proportional to $\psi^N(1 - \psi)^{M-N}$ for a model with N individuals in the population and include a hierarchical $Be(1, 1)$ prior distribution for ψ .

If we implement this scheme using a TD Gibbs sampler (e.g. Link and Barker 2010, pg. 146), where in each iteration we choose between pairs of these models that correspond to $w_i = 0$ or $w_i = 1$, then the resulting TD algorithm has the same full conditional distributions as the DA algorithm described in Royle et al. (2007). In this sense, the TD algorithm described above is exactly the same as the equivalent DA algorithm and any distinctions can be thought of as semantic.

Any difference that does exist between the two approaches lies in the interpretation that can be placed on supplemental variables. In the fixed-dimension DA approach, the $M - N$ supplemental capture probabilities correspond to phantom individuals that exist in a sense, but only in the ‘otherworld’ that is described by the superpopulation. Of necessity, these capture probabilities have the same prior as any other capture probability in the model. In the TD approach described above, as well as the standard TD approach to M_h , the supplemental variables have only a dimension-matching role. Their distribution is chosen solely to ensure an efficient Gibbs sampler and they have no interpretation as parameters within the context of the model.

It is not surprising that the two methods correspond to the same Gibbs sampler, since a crucial aspect of the RJMCMC algorithm as laid out by Green (1995) is the use of auxiliary variables to match the dimension of the various models. In this case, we can think of RJMCMC as using DA to match dimensions. So long as we choose these auxiliary variables appropriately, as we did above, we are able to make the DA scheme of Royle et al. (2007) correspond to a RJMCMC scheme by only considering RJMCMC algorithms that use identity bijections between models.

4 Discussion

While the super-population model as implemented by Royle et al. (2007) is simple and appropriate for most analyses we emphasize that extensions may be difficult or problematic to incorporate in this modeling framework. The BUGS code we have provided for the standard approach offers an easy solution if other hierarchical priors are of interest for N .

Extensions to include changes such as relaxing the assumption of error-free tag reading (e.g., Wright et al. 2009) are easy to include in the standard approach provided an explicit reversible-jump step is included in the Gibbs sampling algorithm. It is also important in this context that we do not condition on one ordering of the \mathbf{x} (or equivalently \mathbf{y}_i = the number of times individual i is caught) as is done in the BUGS code of Royle et al. (2007) as well as the BUGS code in Figure 1. Irrespective of the model we choose to use, care needs to be taken to ensure the combinatorial term is correct for the given model. For the full likelihood approach, this means that we include the full combinatorial term, as shown in (3), in the model when we sample from the full conditional distributions for all unknowns, including the uncertain tags (see Wright et al. (2009) for an example). It is less clear how we would proceed in the superpopulation approach in the presence of tag error.

Being able to specify the model without having to explicitly include the combinatorial terms is one of the advantages of the superpopulation approach. However, as we have shown, while correct for the standard model, conditioning on $W_1 = 1, \dots, W_n = 1$ results in an marginal likelihood that has the incorrect combinatorial term if there is error in tag reading. In order to obtain valid inference in this setting, additional adjustments to the superpopulation model will be required to ensure that the marginal likelihood is correct.

We emphasise that although we have included an upper limit M on N in our BUGS code, this is not required for the general TD algorithm. Also, during general implementation we only need sample values for auxiliary variables that are involved in

the change of dimension from N to N^* rather than the $M - n$ sampled by the BUGS code. This can result in substantial gains in efficiency when M is much larger than a typical value for N , for example, when there is skew in the posterior for N .

We believe that the main advantage of the super-population approach as advocated by Royle et al. (2007) is the fast and easy implementation of the model in BUGS. However, an unappealing aspect of this approach is that N is not explicit in the likelihood and its use appears to be restricted to implied priors for N that are based on binomial mixture models. BUGS code for fast and easy implementation of the standard model can also be written as we have shown.

Although we have focused here on the mark-recapture model M_h , the insights we offer are not restricted to M_h but extend to other closed and open population capture-recapture models, occupancy modeling, distance sampling, and any other problem that involves inference about an unobserved number of draws from a series of categorical distributions each with a distinct probability vector.

References

- Burnham, K. P. and Overton, W. S. (1978), “Estimation of the size of a closed population when capture probabilities vary among animals,” *Biometrika*, 65, 625–633.
- Carlin, B. and Chib, S. (1995), “Bayesian model choice via Markov chain Monte Carlo methods,” *Journal of the Royal Statistical Society. Series B*, 57, 473–484.
- Durban, J. W. and Elston, D. A. (2005), “Mark-Recapture With Occasion and Individual Effects: Abundance Estimation Through Bayesian Model Selection in a Fixed Dimensional Parameter Space,” *Journal of Agricultural, Biological, and Environmental Statistics*, 10, 291–305.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, 2nd ed.

- Green, P. J. (1995), “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination,” *Biometrika*, 82, 711 – 732.
- King, R. and Brooks, S. P. (2008), “On the Bayesian Estimation of a Closed Population Size in the Presence of Heterogeneity and Model Uncertainty,” *Biometrics*, 64, 816–824.
- Link, W. A. (2003), “Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities,” *Biometrics*, 59, 1123–1130.
- Link, W. A. and Barker, R. J. (2010), *Bayesian Inference with ecological applications*, Academic Press.
- Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000), “WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility,” *Statistics and Computing*, 10, 325–337.
- Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978), “Statistical inference from capture data on closed animal populations.” *Wildlife Monographs*, 62, 1–135.
- Pledger, S. (2000), “Unified maximum likelihood estimates for closed capture-recapture models using mixtures,” *Biometrics*, 56, 434–442.
- Plummer, M. (2003), “JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling,” in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March*, pp. 20–22.
- Royle, J. and Dorazio, R. (2008), *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*, Academic Press.
- Royle, J. A., Dorazio, R. M., and Link, W. A. (2007), “Analysis of multinomial models with unknown index using data augmentation,” *Journal of Computational and Graphical Statistics*, 16, 67–85.

- Schofield, M. R. and Barker, R. J. (2008), “A unified capture-recapture framework,” *Journal of Agricultural, Biological and Environmental Statistics*, 13, 458–477.
- (2010), “Full Open Population Capture-Recapture Models with Individual Covariates,” *Journal of Agricultural, Biological and Environmental Statistics*, Submitted.
- Tanner, M. A. and Wong, W. H. (1987), “The Calculation of Posterior Distributions by Data Augmentation (with discussion),” *Journal of the American Statistical Association*, 82, 529–550.
- Wright, J., Barker, R., Schofield, M., Frantz, A., Byrom, A., and Gleeson, D. (2009), “Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples,” *Biometrics*, 65, 833–840.

Definition	
k	the number of occasions on which the population is sampled.
N	population size.
M	superpopulation size ($M \geq N$).
\mathbf{P}	an $N \times 1$ random vector of capture probabilities with a realisation denoted by \mathbf{p} .
\mathcal{P}	the sample space for \mathbf{P} .
\mathbf{X}	a random capture history matrix of dimension $N \times k$ with X_{ij} the indicator for capture of individual i ($i = 1, \dots, N$) in sample j ($j = 1, \dots, k$). We use \mathbf{x} to denote a realisation of \mathbf{X} and \mathbf{x}^{obs} a value observed after sampling; \mathbf{x}^{obs} includes $N - n$ rows of zeros which are treated as observed conditional on N .
\mathcal{S}	the sample space for \mathbf{X} .
\mathbf{W}	a random vector with value 1 if the i th member of the superpopulation is one of the N individuals that are in the study population. We denote a realisation of \mathbf{W} by \mathbf{w} .
\mathcal{W}	the sample space for \mathbf{W} .
z_h	the number of individuals with capture history $h \in \mathcal{H}$, where $\mathcal{H} = \{11 \dots 11, 11 \dots 10, \dots, 00 \dots 00\}$. We also refer to the null history $00 \dots 00$ as $\mathbf{0}$.

Table 1: Notation

```

1. model {
2.   mu ~ dlogis(0,1)
3.   t ~ dt(0,1,2)
4.   tau<-abs(t*5)
5.   for(i in 1:M){
6.     logit(p[i]) <- lp[i]
7.     lp[i] ~ dnorm(mu,tau)
8.     y[i] ~ dbin(pi[i],k)
9.     pi[i] <- p[i]*w[i]
10.    w[i] <- step(N-i)
11.  }
12. N ~ dcat(pee[1:M])
13. n ~ dbin(0.00001,N)
14. }

```

Explanation of the BUGS code:

Line 2–4 & 6–7 Distribution $f(p|\theta_p)$, including prior distributions on $\theta_p = (\mu, \tau)$.

Line 8–10 Modeling the observed capture histories – $y[i]$ is the number of times individual i was caught during k samples. We could model the capture indicators structure $x[i, j]$ if we prefer.

Line 12 Specifying $f(N)$. Note that the starting value for $N > n$. Here we chose to use a discrete uniform prior on N so the values `pee[i]` are specified as data and are all set to $\frac{1}{M}$. Note that (i) there are many different ways we can parameterize this distribution and that this may affect the mixing speed as well as the time taken to run the code.

Line 13 Including the term $\propto \frac{N!}{(N-n)!}$ using the approximation of Durban and Elston (2005) who treat n as binomially distributed with index N and success probability $\epsilon \approx 0$. We assume that tags are correctly read and so do not include $\prod y_h!$. Another approach to including the normalizing constant is to include a variable `zero=0` and then substitute line 13 with two lines: (i) `zeros~dpois(lam)` and (ii) `lam <- -(logfact(N)-logfact(N-n)-logfact(M))`.

Note that this model statement is appropriate for both BUGS and JAGS. We provide alternative model statements and the data and initial values required for both BUGS and JAGS for each of these modeling statements at www.maramatanga.com.

Figure 1: BUGS Code For Model Mh Using RJMCMC

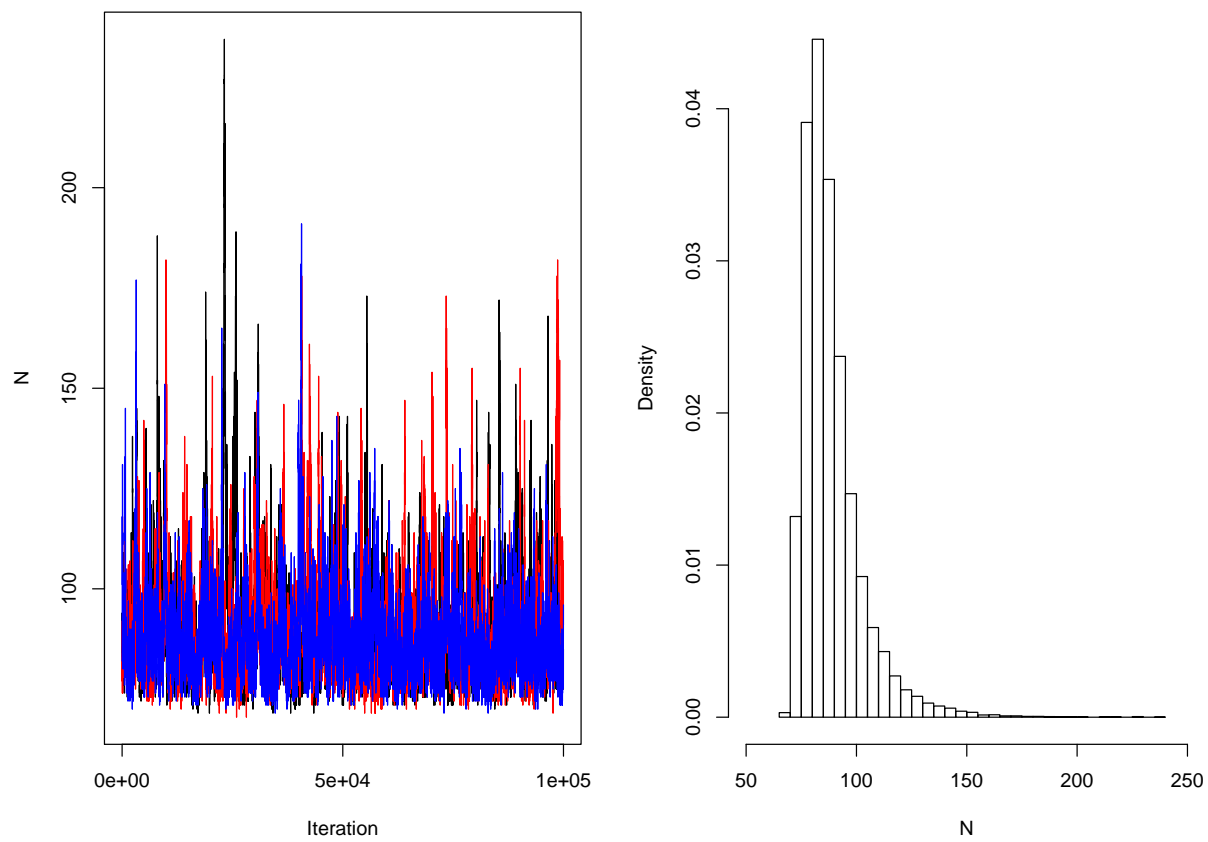


Figure 2: Traceplot and Posterior Density for N for the snowshoe hare data fitted in JAGS using the code in figure 1.